Supplementary Materials (SM):

Manuscript title: Coevolution in defining the functional specificity.

Authors: Saikat Chakrabarti and Anna R. Panchenko.

Table SM1: Description of the dataset.

Code	Description	No. of subgroup	No. of Family member	Representative PDB structure code	RMSD*	LHM*	No. of subsite	Alignment length	Avg. sequence identity (%)	Alignment location of subsites
cd00120	MADS: MCM1, Agamous, Deficiens, and SRF box family.	2	90	1EGW_A	0.9	0.0	3 [1,2]	1108	12	479,487 490
cd00264	Bactericidal permeability-increasing protein, lipopolysaccharide-binding protein and cholesteryl ester transfer protein domains	2	31	1BP1_1	--	--	3 [3-6]	831	8	317, 323,330
cd00333	Major intrinsic protein (MIP) family	2	27	1FX8_A	2.2	7.4	12 [7-10]	1118	20	358,362,418,423,489,624,628,638,639,645,649, 675
cd00363	Phosphofructokinase	2	11	1PFK_A	0.9	0.3	6 [11]	1101	35	224,291,294,295,375,559
cd00365	Hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase	2	30	1DQA	1.8	3.3	10 [12]	1151	24	731,733,734,737,945, 1054, 1057, 1058, 1070, 1074
cd00423	Pterin binding enzymes	2	33	1AJ0	2.0	2.4	4 [13-15]	1394	16	690,691,739,740

cd00985	Maf_Ham1 family	2	180	2MJP_A	2.1	4.9	3 [16-17]	1051	17	247,249,258
Gprotein	G protein alpha subunit	11	105	1FQJ	0.8	0.2	7 [7,8,18]	310	47	16,115,206,214,220,222,306
GST	Glutathione S-transferase family	11	107	2GST	2.5	3.1	9 [7, 19]	330	20	9,10,20,22,23,33,34,126,140
LacI	LacI/PurR family	15	54	1WET	0.2	0.0	12 [7,8,18,20]	340	27	14,15,49,54,84,97,113,121,122,145146,159,220,248
Ricin	RICIN domain family	3	47	1ISY	1.3	3.9	21 [21-22]	135	37	1,15,30,32,38,40,49,50,55,59,63,64,65,109,111,112,119,130,132,133,134
CBM9	Family 9 carbohydrate-binding module	2	19	1I82	0.1	0.0	7 [23]	196	37	73,79,100,102,157,179,182

***** Root mean square deviation (RMSD) and loop Hausdorff similarity metric(LHM) are measures of structural similarity and are averaged over all structures in a given domain family; LHM is used for calculation of structural (dis)similarity within the loop regions[24].

Data collection

MADS:
MCM1, Agamous, Deficiens, and SRF (serum response factor) box family of eukaryotic transcriptional regulators form the MADS family. These proteins bind DNA and exist as hetero and homodimers. This family is composed of 2 main subfamilies: SRF-like/Type I and MEF2-like (myocyte enhancer factor 2)/ Type II as suggested by CDD database[25]. Apart from the extra alpha-2 helix responsible for the dimerization interface in SRF-like/Type I subfamily, there are three other sites that could be important in specificity determination. Two of these sites (alignment columns 487 and 490; see Table SM1) were identified as phosphorylation sites in the MADS_MEF2-like subfamily and were linked to the increased DNA binding affinity[1, 2] . The third subsite is a part of the dimerization interface.

Bactericidal/permeability-increasing protein domain:
Bactericidal permeability-increasing (BPI) proteins bind to and neutralize lipopolysaccharides (LPS) from the outer membrane of Gram-negative bacteria. Apolar pockets, formed mainly by helix-A on the concave surface bind a molecule of phosphatidylcholine, primarily by interacting with its acyl chains. It consists of two domains of similar sizes (N-terminal BPI1 and C-terminal BPI2) that are connected by a proline rich linker of 21 residues (positions 230 to 250).The N-terminal domain of BPI is cationic and retains the bactericidal, LPS binding, and LPS-neutralization activities of the intact protein[3,4]. The COOH-terminal domain is essentially neutral and shows limited LPS-neutralization activity[5]. From the study of Beamer *et al*., 1997[6] we extrapolated three sites (Arg8, Leu14 and Gly21 in BPI1) residing just before and within the helix A and A` (in BPI1 and BPI2, respectively); these sites could be responsible for variable bactericidal binding properties and may be important for subfamily specificity.

Major intrinsic protein (MIP) family:
The major intrinsic protein (MIP) family is a large and diverse family of transmembrane channels containing two major subfamilies with bacterial members: glycerol-transporting channel proteins (GLP) and aquaporins (AQPs), water-transporting channel proteins. For the current study 12 sites were selected as specificity determining sites involving in either pore selectivity, water channel formation in AQPs or interaction with glycerol in GLPs[7- 10].

Phosphofructokinase:
Phosphofructokinase (PFK) catalyzes the phosphorylation of fructose-6-phosphate to fructose-1,6-biphosphate. PFK family contains two subfamilies; ATP and pyrophosphate (PPi) dependent phosphofructokinases. Generally, ATP-PFKs are allosteric homotetramers, and PPi-

PFKs are dimeric and nonallosteric except for plant PPi-PFKs which are allosteric heterotetramers. Six sites that have been suggested to be important in maintaining specific binding to ATP or PPi and they were selected as true positives for this analysis[11].

HMG-CoA reductase:
Hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase (HMGR) is a tightly regulated enzyme, which catalyzes the synthesis of coenzyme A and mevalonate in isoprenoid synthesis. There are two classes of HMGR: class I enzymes which are found predominantly in eukaryotes and contain N-terminal membrane regions and class II enzymes which are found primarily in prokaryotes and are soluble as they lack the membrane region. Human (belongs to class I subfamily) and bacterial HMGR (belongs to class II subfamily) differ in their active site architecture[12,13]. Class I HMGRs generally binds to HMG, HMG-CoA in a NADP dependent reaction while class II HMGRs binds to HMG-COA, mevalonate and NAD. Ten sites were selected as specificity determining at which most differences observed in catalytic and substrate binding properties between the class I and II HMGRs[12,13].

Pterin binding enzymes:
This family includes dihydropteroate synthase (DHPS) subfamily and cobalamin-dependent methyltransferases[24]. Both DHPS and cobalamin-dependent methyltransferases bind to pterin substrates while sulfonamide drugs act as a specific ligand to DHPS. Four sites that could to be important for sulfonamide binding in DHPS[14-16] were considered as specificity determining sites.

Maf_Ham1 family:
Maf_Ham1 domain family contains two subfamilies, Ham1 and Maf. A Ham-related protein from *Methanococcus jannaschii* is a novel NTPase that has been shown to hydrolyze nonstandard nucleotides, such as hypoxanthine/xanthine NTP, but not standard nucleotides. Maf, a nucleotide binding protein, has been implicated in inhibition of septum formation in eukaryotes, bacteria and archaea. Three conserved residues could be important for binding to different nucleotides [*e.g.*, 2'-Deoxyuridine 5'-Triphosphate (dUTP) and Xanthosine 5'-Triphosphate (XTP) for Maf and Ham1, respectively] and therefore can be regarded as specificity determining sites[17, 18].

G protein alpha subunit family:
G_α subunits of G protein can be divided into four main subtypes where each of the classes performs different biological functions through specific interactions with the effectors [*e.g.* cyclic GMP phosophodiesterase (PDE)] and regulators [*e.g.* Regulator of G protein signaling (RGS) domains]. Multiple sequence alignment were obtained from Pei *et al*., 2006[19], where the main four subtypes are further divided into 11 subfamilies depending either on the taxonomy (like, plant, animal, fungal G proteins) and type of functions involved (e.g. stimulation G_s; inhibition G_i, etc). Sites that were predicted as being specificity determining by both SPEL[19] and SDP-pred[7, 8] methods and were also found to be spatially proximal to the specific effectors or regulators were considered true positives in our analysis.

Glutathione S-transferase family:
Glutathione S-transferase (GST) enzymes function to detoxify a wide variety of xenobiotic substrates with reactive electrophilic groups by conjugation to the tri-peptide glutathione (GSH). GST enzymes have been extensively characterized and are grouped according to a robust

classification system that is based on a variety of criteria including primary structure, immunoblotting, kinetic properties, inhibitor sensitivity, tertiary structure, and quaternary structure[20]. Multiple sequence alignment of the GST family and the specificity determinant sites were obtained from Pei *et al*., 2006[19].

LacI/PurR family:

The LacI/PurR family of transcription factors is regulated by small molecules, such as sugars and nucleotides. In addition to available experimental and structural information, the LacI/PurR family has been widely used by researchers for prediction of specificity determining sites. This family contains 15 specificity groups: AraR, KdgR, CcpA, DegA, YjmH, RbsR, PurR, CytR, GalSR, AscG, LacI, TreR, GntR, IdnR, and FruR. Generally, researchers[7,8,19,21] have identified specificity determining sites through examination of possible contacts between ligand molecules (effector and DNA) and amino acid residues or between amino acid residues of different subunits.

RICIN domain family:

A single RICIN domain can be divided into three structural domains and three corresponding domain subfamilies of approximately 40 amino acids in length that have evolved from an ancient galactose binding peptides[22]. The first domain subfamily possesses two carbohydrate binding sites and a peptide binding region (group II), the second domain subfamily contains the N-terminal carbohydrate binding region (group I) and the third one covers the C-terminal carbohydrate binding region (group III)[7]. Multiple alignments and the information regarding the sites that could be specific to each domain subfamily were obtained from Pils *et al*., 2005[23].

Family 9 carbohydrate-binding module:

The alignment of carbohydrate-binding module family 9 (CBM9) containing approximately 19 putative sequences from 11 organisms was obtained from Notenboom *et al*., 2001[24]. CBM9 family contains two subfamilies: 9a and 9b. Subfamily-9a comprises the N-terminal part of tandem CBM9 modules. The subfamily-9a module suggests lack of carbohydrate-binding function compared to the other subfamily-9b which binds to amorphous and crystalline cellulose and different soluble di- and monosaccharide[24].

Table SM2: List of CDD families and their functionally important sites (FIS).

Code	Description	No. of Family member	Alignment length	Avg. sequence identity (%)	Number of functionally important sites
cd00120	MADS: MCM1, Agamous, Deficiens, and SRF box family.	90	1108	12	35
cd00264	Bactericidal permeability-increasing protein, lipopolysaccharide-binding protein and cholesteryl ester transfer protein domains	31	831	8	19
cd00333	Major intrinsic protein (MIP) family	27	1118	20	12
cd00363	Phosphofructokinase	11	1101	35	37
cd00365	Hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase	30	1151	24	59
cd00423	Pterin binding enzymes	33	1394	16	14
cd00985	Maf_Ham1 family	180	1051	17	5

Figure SM1:

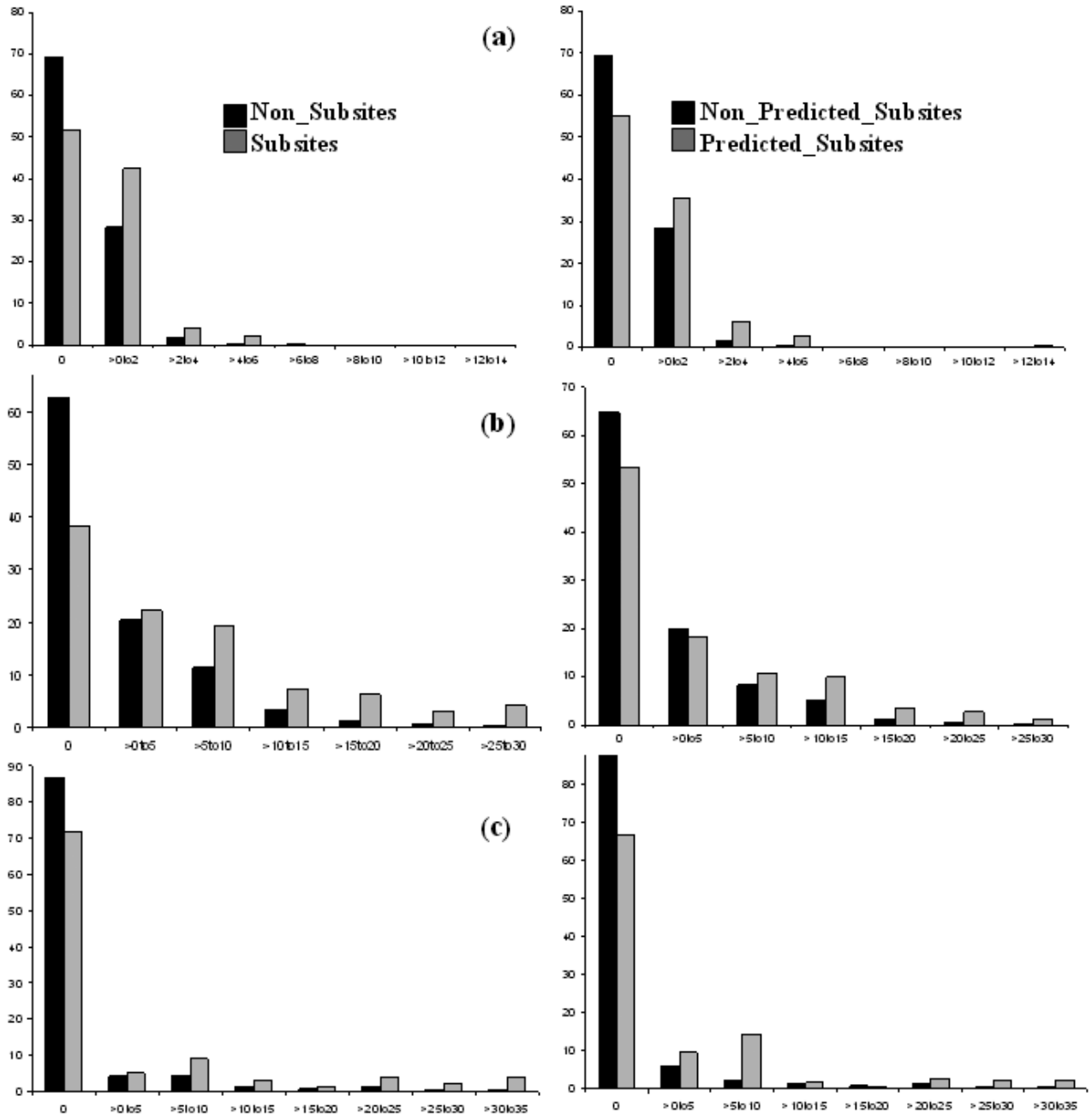

Figure SM1: Fraction of coevolution (FC) for actual and predicted subsites. Fraction of coevolution (shown in % scale bins) for each site was calculated as the number of suggested coevolved site pairs divided by all possible site pair combinations for a given site with all other sites. Different coevolution prediction algorithms were applied to calculate FC within our dataset. a) FC

calculated using the MIp algorithm[26] for actual subsites and non-subsites (upper left panel) and for top SPEER[27] predicted subsites (upper right panel); b) FC calculated using the OMES[28,29] algorithm for actual subsites and non-subsites (middle left panel) and for top SPEER predicted subsites (middle right panel); c) FC calculated using the McBASC[30-32] algorithm for actual subsites and non-subsites (lower left panel) and for SPEER top predicted subsites (lower right panel).

Figure SM2:

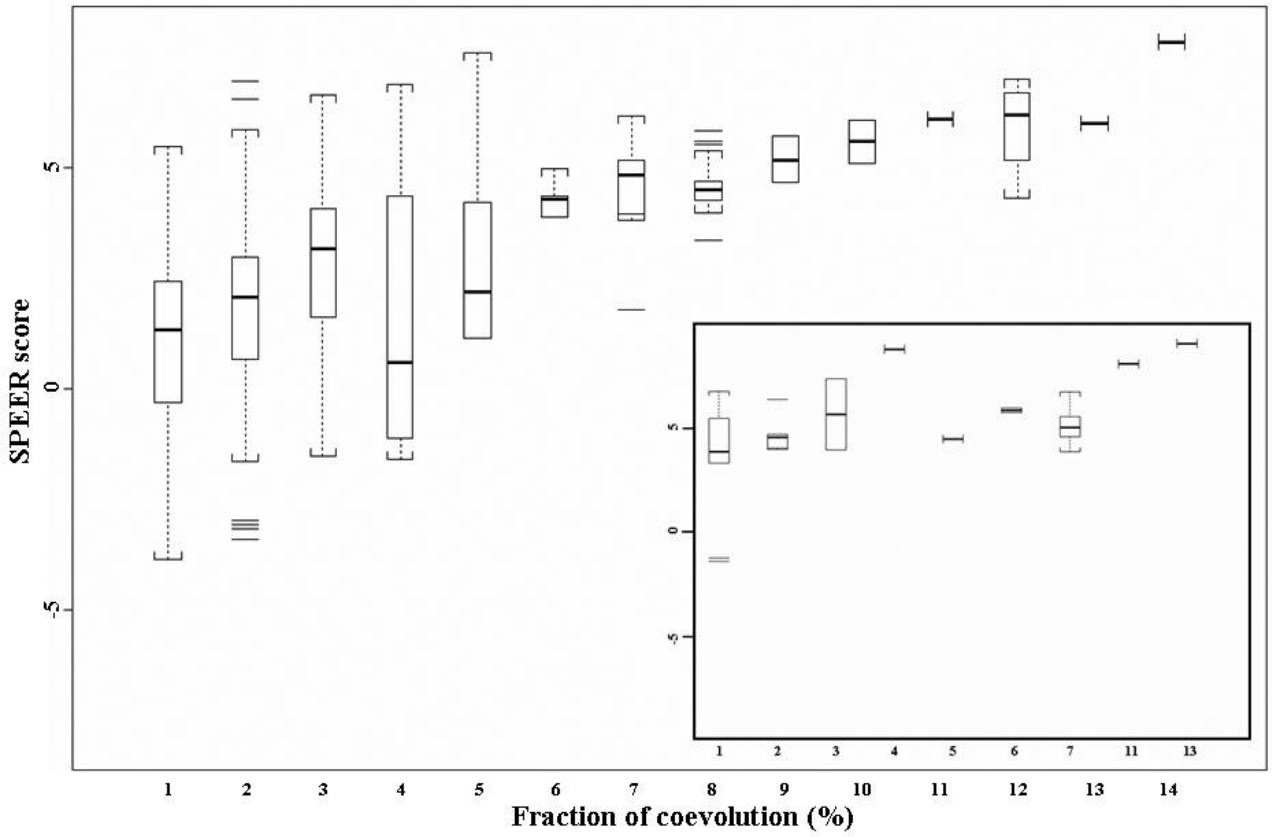

Figure SM2: Correlation of SPEER score with fraction of coevolution for a given site. Fraction of coevolution (shown in % scale) for each site is plotted against the SPEER[27] score. The inset shows the fraction of evolution for actual subsites only. The central line in each box shows the median value, the upper and lower boundaries of individual box show the upper and lower quartiles, and the vertical lines extent to a value of 1.5 times the inter quartile range. Outlier values are shown outside the whiskers.

Figure SM3:

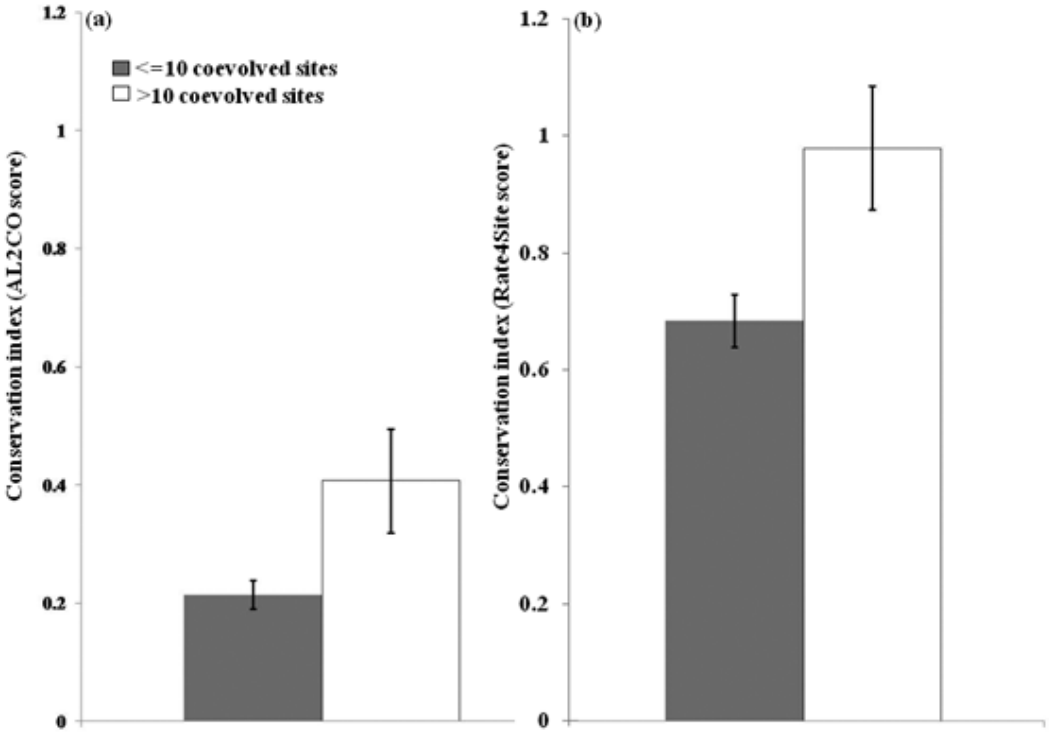

Figure SM3: Correlation of evolutionary conservation with the fraction of coevolution for subsites. Fraction of coevolution for each subsite is plotted against evolutionary conservation [evolutionary conservation scores were calculated by the a) AL2CO[35] and b) Rate4Site[36] programs] where higher values indicate higher conservation. For comparison purpose, Rate4Site scores are projected on a reverse scale. Each box shows the mean value and standard error of fraction coevolution for sites with less that 10 (grey box) and more than 10 coevolutionary connections (open box).

Figure SM4:

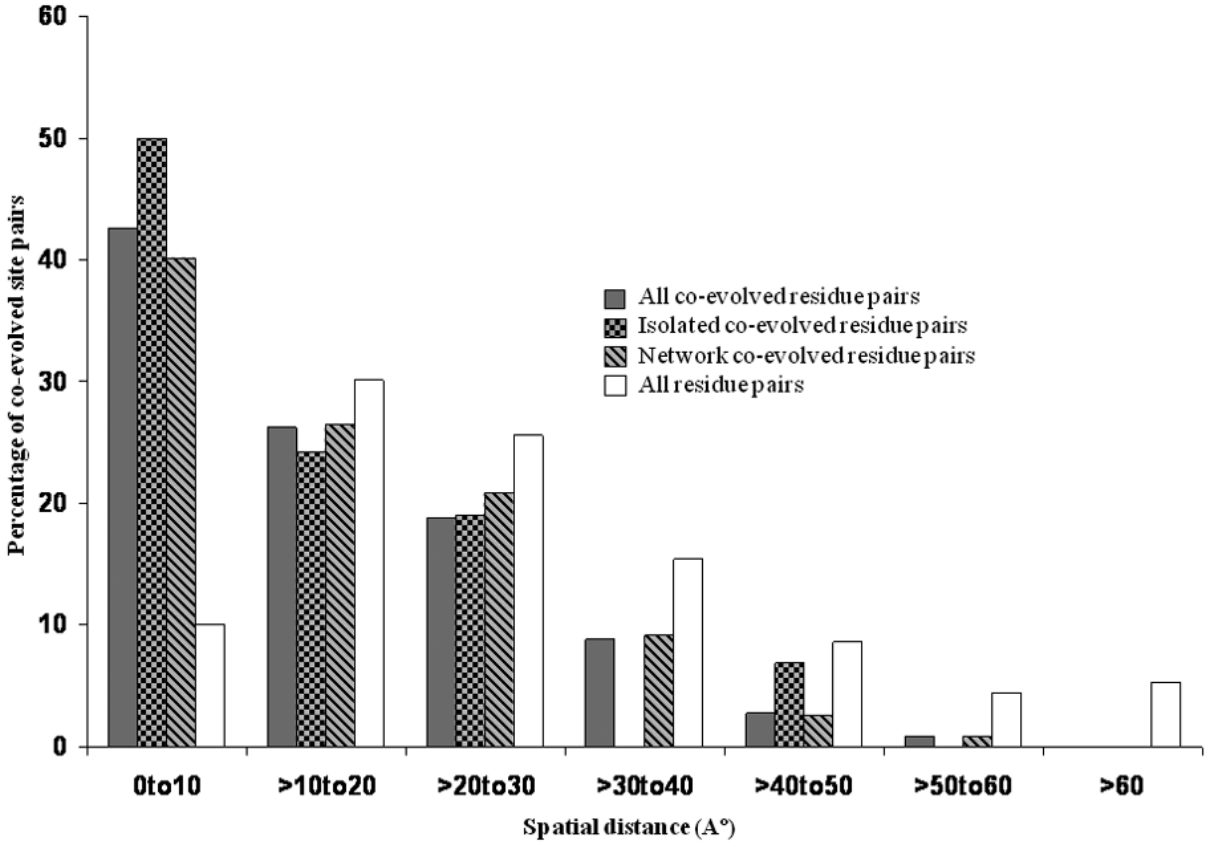

Figure SM4: Spatial distance analysis among all coevolved sites. Spatial distances among all coevolved site pairs (grey bars), isolated coevolved site pairs (grey squared bars), network coevolved site pairs (grey striped bars) and all background residue pairs (open bars) are plotted into bins. Isolated coevolved site pairs[33] were defined as those site pairs that are not coevolved with any other sites whereas site pairs that are also coevolved with other sites forming a network were termed as network coevolved site pairs[34]. The spatial distance distribution of all site pairs within the representative structures are shown in open bars.

Figure SM5:

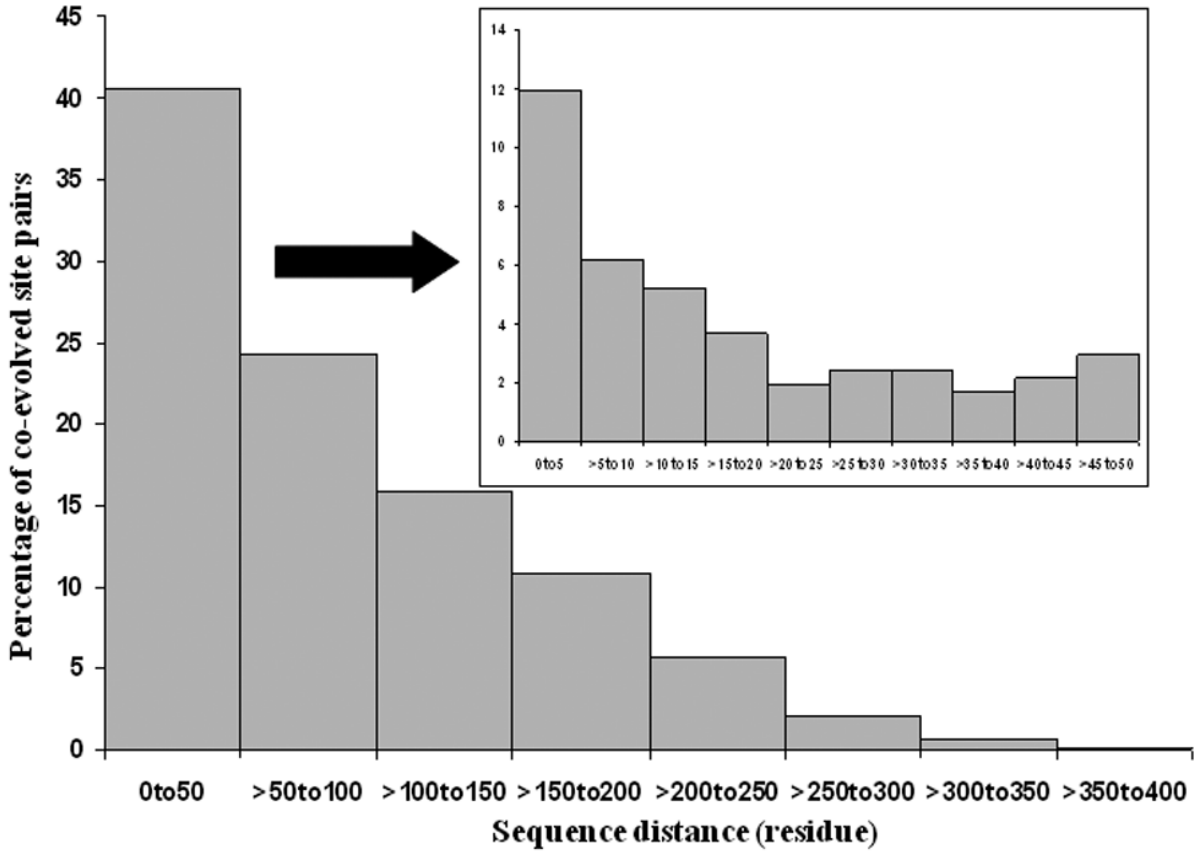

Figure SM5: Sequence distance analysis among all coevolved sites. Sequence distances among all coevolved site pairs were plotted in bins. The inset shows a blow out of distance distribution of coevolved site pairs that are within 50 residues apart.

SM Reference:

1. Santelli E, Richmond TJ. Crystal structure of MEF2A core bound to DNA at 1.5A resolution. J Mol Biol 2000; 297: 437-449.
2. Tan S, Richmond TJ. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature 1998; 391: 660-666.
3. Ooi CE, Weiss J, Elsbach P, Frangione B, Mannion B. A 25-kDa NH2-terminal fragment carries all the antibacterial activities of the human neutrophil 60-kDa bactericidal/permeability-increasing protein. J Biol Chem 1987; 262: 14891-14894.
4. Ooi CE, Weiss J, Doerfler ME, Elsbach P. Endotoxin-neutralizing properties of the 25 kD N-terminal fragment and a newly isolated 30 kD C-terminal fragment of the 55-60 kD bactericidal/permeability-increasing protein of human neutrophils. J Exp Med 1991; 174: 649-655.
5. Abrahamson SL, Wu HM, Williams RE, Der K, Ottah N, Little R, Gazzano-Santoro, H, Theofan G, Bauer R, Leigh S, Orme A, Horwitz AH, Carroll SF, Dedrick RL. Biochemical characterization of recombinant fusions of lipopolysaccharide binding protein and bactericidal/permeability-increasing protein. Implications in biological activity. J Biol Chem 1997; 272: 2149-2155.
6. Beamer LJ, Carroll SF, Eisenberg D. Crystal Structure of Human BPI and Two Bound Phospholipids at 2.4 Angstrom Resolution. Science 1997; 276: 1861-1864.
7. Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. Nucleic Acids Res 2004; 32: W424-428.
8. Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci 2004; 13: 443-456.
9. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, Krucinski J, Stroud RM. Structure of a glycerol-conducting channel and the basis for its selectivity. Science 2000; 290: 481–486.
10. Sui H, Han BG, Lee JK, Walian P, Jap BK. Structural basis of water-specific transport through the AQP1 water channel. Nature 2001; 414: 872–878.
11. Moor SA, Ronimus RS, Roberson RS, Morgan HW. The Structure of a Pyrophosphate-Dependent Phosphofructokinase from the Lyme Disease Spirochete *Borrelia burgdorfer.* Structure 2002; 10: 659–671.
12. Bochar DA, Stauffacher CV, Rodwell VW. Sequence Comparisons Reveal Two Classes of 3-Hydroxy-3-methylglutaryl Coenzyme A Reductase. Molecular Genetics and Metabolism 1999; 66: 122–127.
13. Istvan ES. Bacterial and mammalian HMG-CoA reductases: related enzymes with distinct architectures. Curr Opin Struct Biol 2001; 11:746-751.
14. Hampele IC, D'Arcy A, Dale GE, Kostrewa D, Nielsen J, Oefner C, Page MG, Schonfeld HJ, Stuber D, Then RL. Structure and function of the dihydropteroate synthase from *Staphylococcus aureus*. J Mol Biol 1997; 268: 21-30.
15. Achari A, Somers DO, Champness JN, Bryant PK, Rosemond J, Stammers DK. Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. Nat Struct Biol 1997; 4: 490-497.
16. Doukov T, Seravalli J, Stezowski JJ, Ragsdale SW. Crystal structure of a methyltetrahydrofolate- and corrinoiddependent methyltransferase. *Structure* 2000; 8: 817–830
17. Minasov G, Teplova M, Stewart GC, Koonin EV, Anderson WF, Egli M. Functional implications from crystal structures of the conserved *Bacillus subtilis* protein Maf with and without dUTP. Proc Natl Acad Sci U S A 2000; 97: 6328–6333.
18. Hwang KY, Chung JH, Kim SH, Han YS, Cho Y. Structure-based identification of a novel NTPase from *Methanococcus jannaschii.* Nat Struct Biol 1999; 6: 691-696.
19. Pei J, Cai W, Kinch LN, Grishin NV. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics 2006; 22: 164-171.
20. Sheehan D, Meade G, Foley VM, Dowd CA. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochem J* 2001; 360: 1-16.

21. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J Mol Biol 2002; 321: 7-20.
22. Rutenber E, Ready M, Robertus JD. Structure and evolution of ricin B chain. Nature 1987; 326: 624-626.
23. Pils B, Copley RC, Schultz J. Variation in structural location and amino acid conservation of functional sites in protein domain families. BMC Bioinformatics 2005; 6: 210-219.
24. Notenboom V, Boraston AB, Kilburn DG, Rose DR. Crystal Structures of the Family 9 Carbohydrate-Binding Module from *Thermotoga maritima* Xylanase 10A in Native and Ligand-Bound Forms. Biochemistry 2001; 40: 6248-6256.
25. Marchler-Bauer A, Anderson JB, Derbyshire MK, Deweese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD. *et al*. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007; 35: D237-240.
26. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008; 24: 333-340.
27. Chakrabarti S, Bryant SH, Panchenko AR. Functional specificity lies within the properties and evolutionary changes of amino acids. J Mol Biol. 2007; 373: 801-810.
28. Larson SM, Di Nardo AA, Davidson AR. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions J Mol Biol. 2000; 303: 433-446.
29. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations Proteins 2002; 48: 611-617.
30. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994; 18: 309-317.
31. Olmea O, Rost B Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol. 1999; 293: 1221-1239.
32. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments Proteins 2004; 56: 211-221.
33. Gloor GB, Martin LC, Wahl LM. Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005; *44*: 7156-7165.
34. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics 2005; 21: 4116-4124.
35. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001; 8:700-712.
36. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002; 18 Suppl 1: S71-77.